Matching structure and the cultural transmission of social norms

Mengel, Friederike

Empfohlene Zitierung / Suggested Citation:
Mengel, F. (2008). Matching structure and the cultural transmission of social norms. *Journal of Economic Behavior & Organization*, *67*(3-4), 608-623. https://doi.org/10.1016/j.jebo.2008.01.001

Title: Matching Structure and the Cultural Transmission of Social Norms

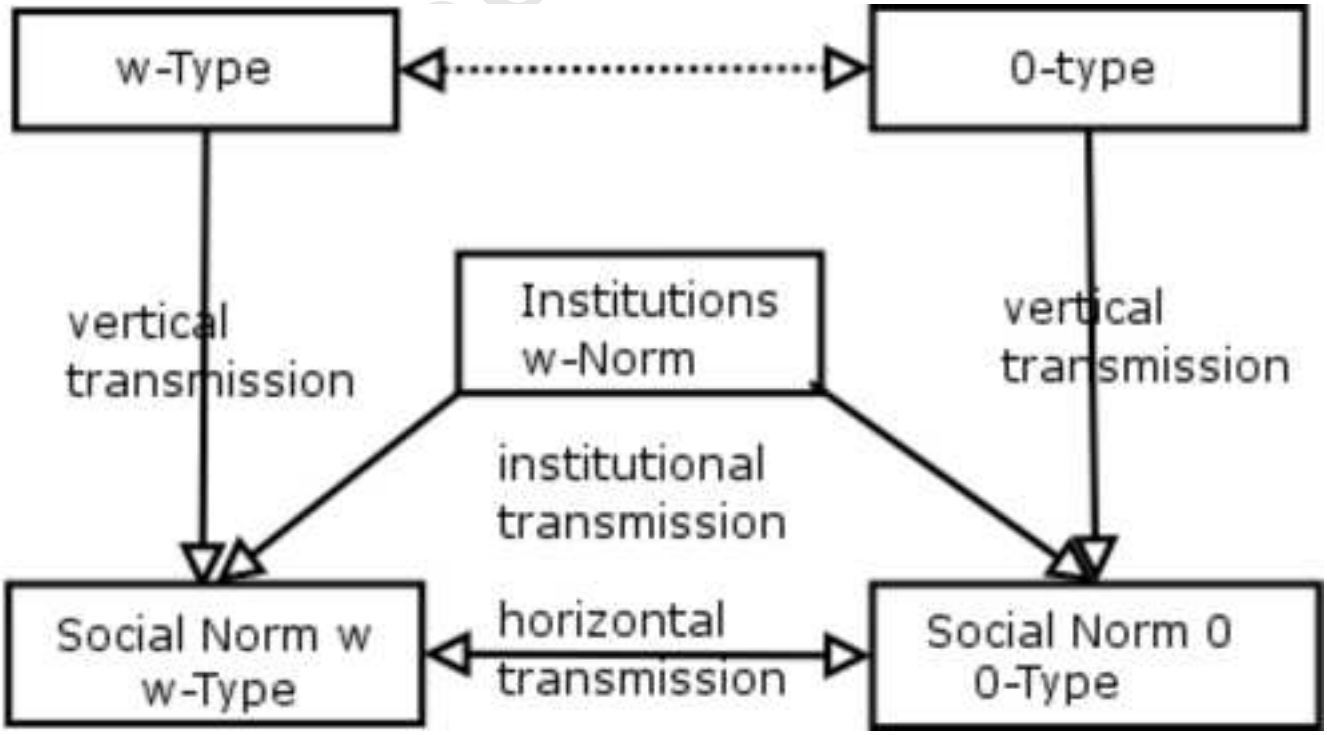
Author: Friederike Mengel

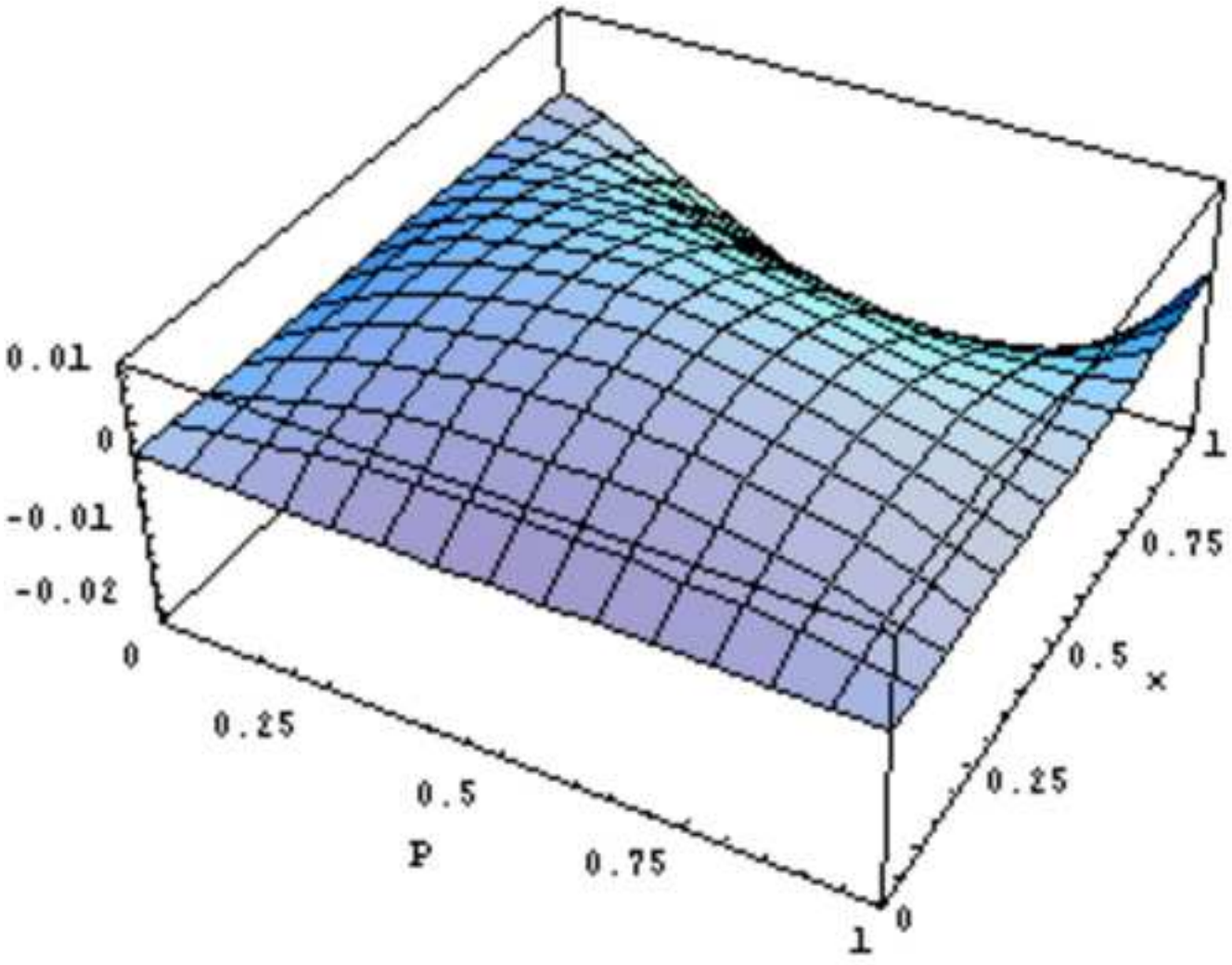
PII: S0167-2681(08)00010-3
DOI: doi:10.1016/j.jebo.2008.01.001
Reference: JEBO 2157

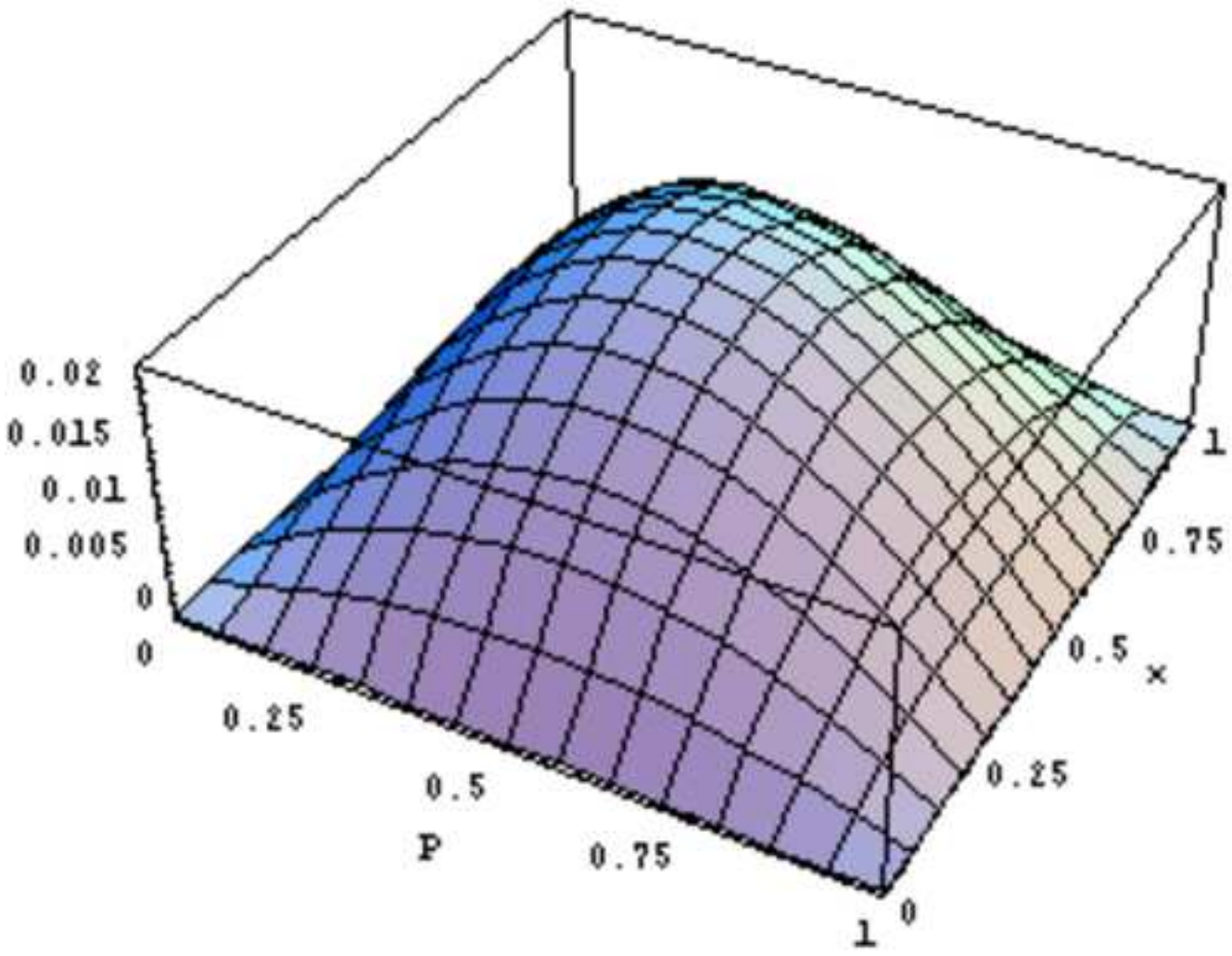
To appear in: *Journal of Economic Behavior & Organization*

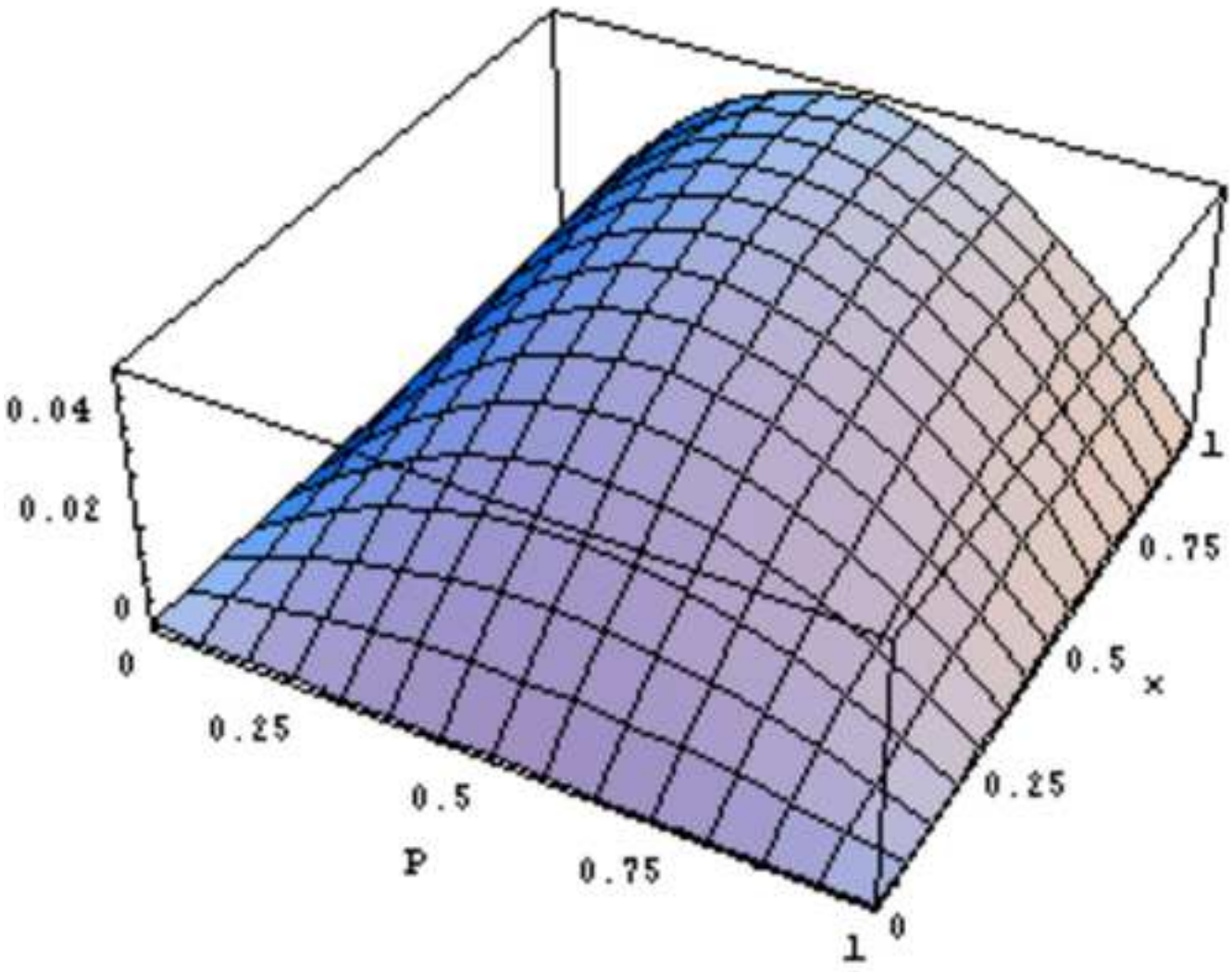
Received date: 11-2-2006
Revised date: 29-12-2007
Accepted date: 7-1-2008

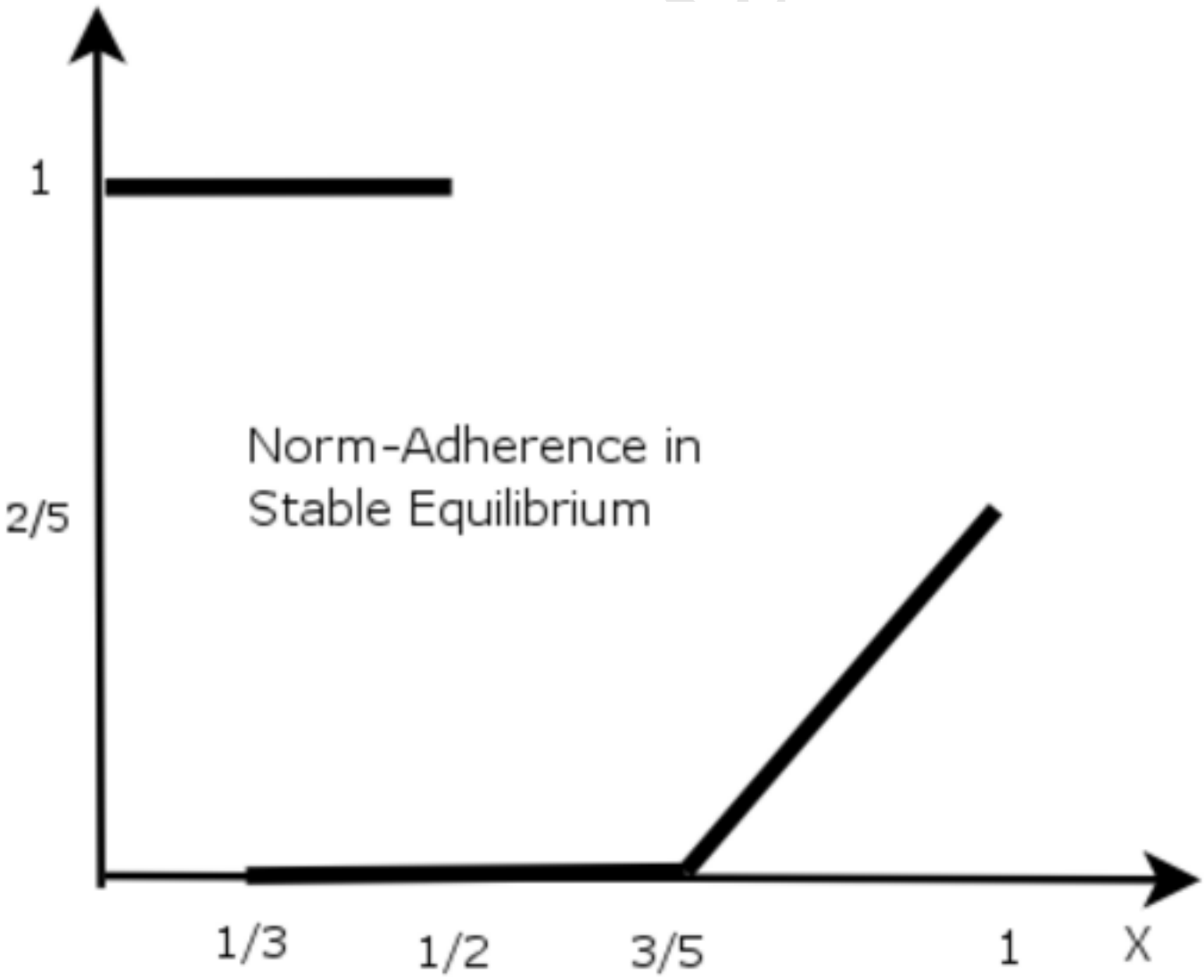
Please cite this article as: Mengel, F., Matching Structure and the Cultural Transmission of Social Norms, *Journal of Economic Behavior and Organization* (2007), doi:10.1016/j.jebo.2008.01.001

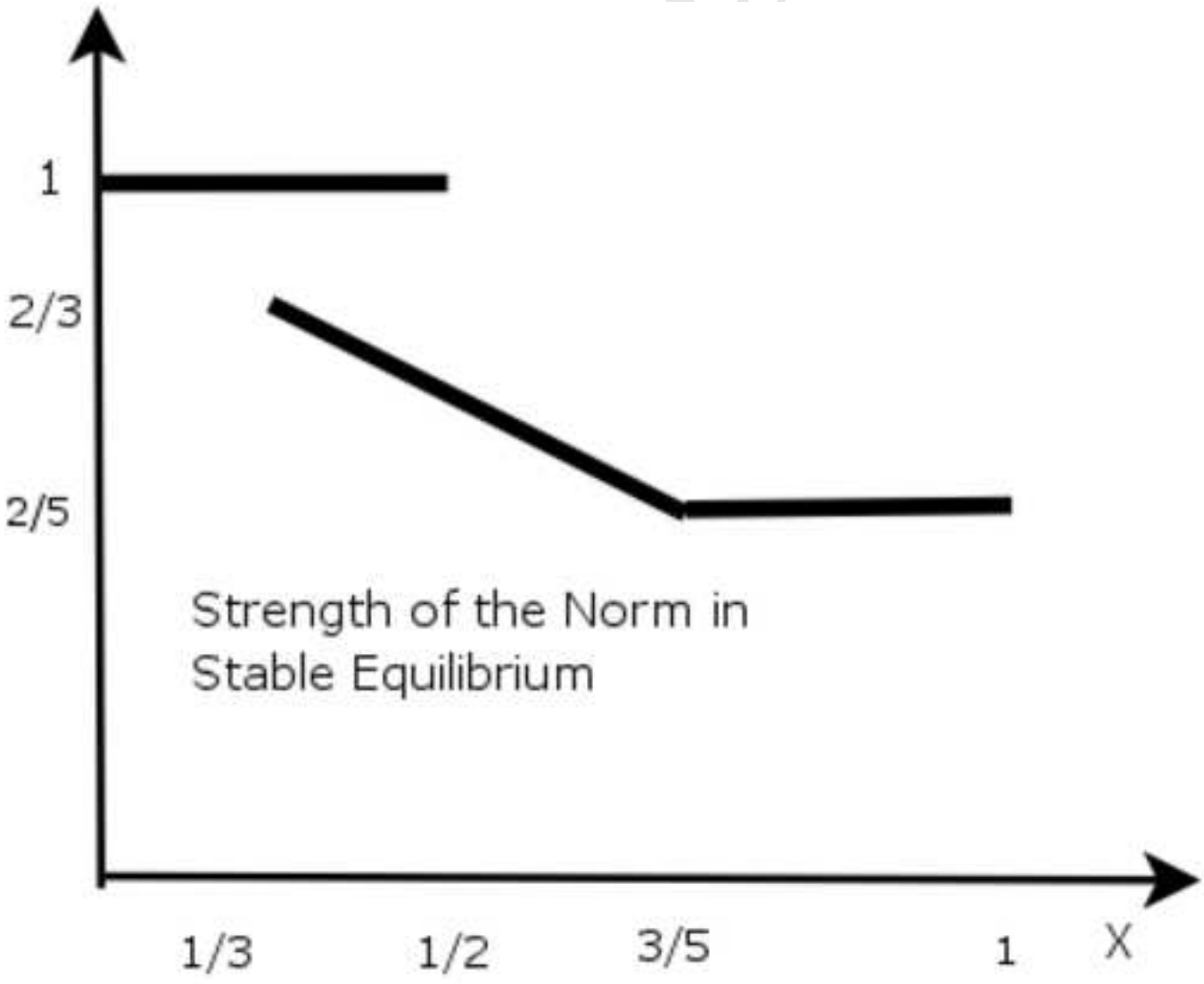












Matching Structure and the Cultural Transmission of Social Norms*

Friederike Mengel†
University of Alicante

First Version: February 2006. This Version: March 2007.

Abstract

We present and study a model of cultural transmission of social norms in a setting where agents are repeatedly matched to play a one-shot interaction prisoners' dilemma. There are two types of agents in the society: some that adhere to a social norm of cooperation and some that don't. Limited integration of these two types can bias the matching structure in the sense that types interact with increased probability among themselves. In contrast to many standard evolutionary approaches, we find that cooperation often survives in the long-run. Specifically we find that while high degrees of separation are needed to protect strict norms for cooperation, norms of intermediate strength can survive in a variety of settings. Endogenizing norm strength, we find two scenarios in which pro-social norms survive. One is a rigid society in which separation leads to equilibria with strict norms for cooperation, and one is an integrated society where equilibria display heterogeneity of types and norms of intermediate strength. Furthermore integration and cooperation are not linked in a monotone way. *JEL-Classification: C70, C73, Z13.*

*I thank Fernando Vega Redondo, Christian Traxler as well as two anonymous referees for very valuable suggestions. I also wish to thank participants at seminars and conferences in A Coruña, Amsterdam, Bayreuth (VFS 2006), Istanbul (SCW 2006), Jena, Karlsruhe, Köln, Valencia and Vienna (ESEM 2006) for their comments. Part of this research was conducted while I was staying at the Max-Planck Institute of Economics in Jena. I thank the Institute for its hospitality. An earlier version of this paper was circulated under the name "A Model of Immigration, Integration and Cultural Transmission of Social Norms".

†*Departamento de Fundamentos del Análisis Economico,* Universidad de Alicante, 03080 Alacant, Spain. *E-mail*: friederike@merlin.fae.ua.es

1 Introduction

In this paper we examine the question of whether pro-social norms can persist in a society if some agents have not internalized these norms.[1] Such agents can appear in a society for example through immigration or through a failure of society to transmit its social norms to part of the population. More precisely we consider a setting where agents are matched to play a one-shot interaction prisoners' dilemma in a society where there is a social norm for cooperation. There are two types of agents. Some have internalized a social norm for cooperation but some have not. We address the following questions.

- Can a social norm for cooperation persist in a society where some agents have not internalized the norm ?
- How does the answer to the previous question depend on the institutions of society and in particular on the degree of integration of the two types ?

From the standard perspective of a *direct evolutionary approach* the first question has a clear-cut answer: If evolutionary selective forces apply directly to strategies, if the population dynamics is payoff monotonic, and if matching takes places randomly within the whole population, cooperation is never evolutionary stable.[2]

In this paper we deviate from the assumptions of the standard approach in three ways. First the cultural transmission mechanism for social norms we model applies to preferences instead of strategies. Holding their preferences fixed, agents are assumed to act rationally.

Secondly we vary the matching technology assuming that different degrees of integration are reflected in the probability that agents of the same type interact with each other. The degree of integration in our approach has two kinds of effects: short-run effects by changing the incentives of rational players and long-run effects by affecting norm strength and the evolution of preferences.

Thirdly we analyze two cases for norm strength. The strength of some norms can be exogenous (independent of the degree of norm-internalization), but for some it will increase with the level of norm-internalization in the society.

We start by analyzing the case of exogenous norm strength and find that for strict norms, a high level of separation or other form of institutional pressure are needed to have cooperation survive. Norms of intermediate strength on the contrary can survive under a variety of settings. Under some parameter constellations they survive even in fully integrated societies. Endogenous norms persist in two polar scenarios. One of a rigid society in which separation leads to monomorphic equilibria with strict norms for cooperation. Cooperation in this scenario is achieved through rigid population structures that in turn lead to strict norms. In this sense rigidity is self-reinforcing. The second scenario is one of an integrated society with intermediate norms that displays heterogeneity

[1] Pro-social norms are norms that induce agents to act in a way conferring benefits to others at a cost to themselves.

[2] Weibull (1995), Vega-Redondo (1996).

of types in equilibrium. Integration stabilizes a polymorphic equilibrium with norms that are less strict. Thus in contrast to standard direct and indirect evolutionary approaches, the mechanism based on endogenous social norms always produces polymorphic equilibria in fully integrated societies (where matching is random). Furthermore we show that integration and cooperation are not linked in a monotone way.

Our approach is very closely related to the *indirect evolutionary approach* in that, while putting selection pressures on preferences it does not deny that holding their preferences fixed, agents act rationally. Bester and Güth (1998) or Guttman (2003) have studied such mechanisms.

The evolution of pro-social preference traits has also been studied in *cultural evolutionary models* that try to go beyond the pure fitness implications of preferences and (induced) strategies and consider explicitly the process of transmission of traits through the family (vertical transmission), peer-groups (horizontal transmission), or socializing institutions of society (oblique transmission).[3] Gintis (2003) presents a model with exogenous vertical and oblique transmission and an (also exogenous) fitness-disadvantage for agents who have a preference for altruism. His main finding is that in order for the altruistic preference to survive the level of oblique transmission has to be sufficiently high. Henrich and Boyd (2001) consider a model in which norms are transmitted through social learning. In their model pro-social norms are stable because the horizontal transmission process stabilizes punishment of non-adherers.

The *rational socialization* approach to preference formation assumes that altruistic and forward-looking parents deliberately pass preferences on to their children, trying to maximize what they, as parents, see to be the children's future well-being. Bisin et al (2004) present a model of endogenous vertical transmission in which altruistic preferences survive, because minorities have higher incentives to socialize their offspring to their own preferences than majorities do.[4]

Whereas all the previous studies consider only the case of random matching, we parametrize several matching scenarios ranging from full separation (where interact only with their type) to the standard case of random matching. These kind of matching structures have been considered mostly in the biological literature.[5] The case typically studied in this literature corresponds to our case of strict norms.

Our study also differs from the above approaches in focusing explicitly on the role of society and social norms. In this sense it relates to studies of norm-guided behavior in other fields of economics. Lindbeck et al (1999) use a model with endogenous social norms to examine the interaction of monetary incentives and social norms in the welfare state. In Benabou and Tirole (2006), norms with endogenous strength are part of a theory of pro-social behavior. To our

[3] See Cavalli-Sforza and Feldman (1981), Henrich and Boyd (2001), Henrich and Gil-White (2000), Boyd and Richerson (2005) or Henrich (2004).

[4] See also Guttman (2001a, 2001b).

[5] See Hamilton (1964), Price (1970), Myerson et al (1991), Henrich (2004), Boyd and Richerson (1990), Mitteldorf and Wilson (2000) or Bowles and Gintis (1997).

knowledge our study is unique in examining the consequence of endogenous social norms for the evolutionary selection of preference traits.[6]

The paper is organized as follows. In section 2 the model is described. In section 3 we study the equilibria of the model with exogenous norm strength, and in section 4 norm strength is endogenized. Section 5 concludes.

2 The Model

Consider a society consisting of a (unit-mass) continuum of individuals I. Individuals are probabilistically and repeatedly matched in pairs to interact in prisoners' dilemma type of situations for an infinite number of periods.

In the bilateral game each player has two actions available: C and D. The *action set* $Z = \{C, D\}$ is the same for all players $i \in I$. Payoffs from the prisoners' dilemma interaction are given by

	C	D
C	a, a	$0, 1$
D	$1, 0$	d, d

(1)

where $1 > a > d > 0$. It is well known that in this game D is a dominant strategy for both players and consequently the unique equilibrium prediction leads to a payoff of d for both players.

2.1 The Social Norm

Definition 1 (Social Norm) An (internalized) social norm is a code of conduct shared by a society and enforced through internal sanctions, including shame, guilt, embarrassment anxiety and loss of self-esteem. As such, an internalized social norm directly affects the utility function.

It is important to distinguish internalized social norms (which we deal with here) from external norms that are sustained through external punishment or social (dis-) approval.[7] Also note that social norms differ from conventions in that the latter are purely descriptive while the first have normative character. A convention can be seen as a behavioral regularity that helps to solve coordination problems in society.[8]

Definition 2 (Norm-internalization) We say that a person "has internalized a social norm" whenever he or she suffers internal sanctions upon deviating from the norm.

[6] Obviously our study also ties in with other studies of norm-guided behaviour such as Azar (2001), Elster (1989), Nyborg and Rege (2003), Traxler (2005), Young (1998), Cialdini et al (1990), Grasmick and Green (1980), Liu (2003) or Reno et al. (1993) among many others.

[7] This would be an alternative modeling approach. In this case the analysis becomes slightly more complicated but the results do not change qualitatively.

[8] To distinguish these terms is important as they have been used in the literature in different ways. Our definition follows Cialdini et al (1990), Grasmick and Green (1980), Elster (1989), Gintis (2003), Lindbeck et al (1999) or Traxler (2005) among others. Young (1998) for example uses a different terminology.

Note that this implies that a person who has internalized the norm need not necessarily behave in the way the norm prescribes whenever the (material) costs of doing so are too high. For agents who have internalized the social norm the psychological cost w associated with deviating from it is reflected in the payoffs as follows:

	C	D
C	a, a	$0, 1-w$
D	$1-w, 0$	$d-w, d-w$

(2)

We will distinguish between three different strengths of the norm. In particular we will call the social norm *weak* if $w < \min\{1-a, d\}$, i.e. if violation of the norm causes feelings of guilt so weak that they are always outweighed by the material payoff-advantage of defecting (playing D). In this case the two game forms (1) and (2) represent the same strategic context, namely that of a prisoners' dilemma. We will call the norm *intermediate* in the following two cases: if $w \in [1-a, d]$ (2) represents a stag-hunt game, having two symmetric Nash-equilibria in pure strategies where both agents play the same strategy (either C or D) and if $w \in [d, 1-a]$ then (2) represents a chicken game, with two asymmetric Nash-equilibria in pure strategies where one plays C and the other plays D. The unique symmetric Nash-equilibrium in this case is in mixed strategies where each player plays $\frac{w-d}{1-a-d}C \oplus \frac{1-a-w}{1-a-d}D$. Finally we will call a norm *strict* if $w > \max\{1-a, d\}$ (if the internal punishment caused by a norm-violation is so strong that cooperation is a dominant strategy for an agent having internalized this norm).

2.2 The Population Game

Types

Let the *type space* be $T = \{0, w\}$ with typical element τ, where a w- type's payoffs are defined in (2) and a 0- type's payoffs in (1). w- types have internalized the norm and 0- types have not. Agents have incomplete information about each other's type. When choosing an action $z \in Z$ in the bilateral game they estimate the type of their match from the distribution of types in the economy and from their knowledge about the matching technology described below. Let p denote the share of w- types in the population. Obviously then the share of 0- types is $1-p$. A complete description of the population is given by the *population profile* (σ^0, σ^w, p) where $\sigma^\tau = (\sigma^\tau_C, \sigma^\tau_D)$ denotes the distribution of actions among τ-types. σ^τ_C is the share of τ- types that use action C.[9] The population profile is known to all agents at all times.

Matching

Matching takes place in a not - fully - integrated population, where individuals interact with increased probability with individuals of their type. We measure the degree of integration of a society with the parameter $x \in [0, 1]$,

[9]Note that $\sigma = (\sigma^0, \sigma^w)$ can be seen as formally equivalent to a pure strategy in the (bilateral) Bayesian game where σ^0 denotes the (mixed) action a player chooses conditional on being a 0- type and σ^w the action a player chooses conditional on being a w- type.

where $x = 1$ means that the society is fully integrated. In this case matching is random. $x = 0$ means that there is full separation, implying that types interact with probability 1 among themselves and never with agents of another type. More precisely a w- type is matched with probability $(1-p)x$ with a 0-type and with probability $1-(1-p)x$ with another w- type, while a 0- type is matched with probability px with a w- type and with probability $(1-px)$ with another 0- type.

Material Payoffs and Utility

For any fixed distribution of types p and degree of separation x denote $\Pi^\tau(z,\sigma)$ the *expected material payoffs* of a type τ agent when choosing action z in a population that plays $\sigma = (\sigma^0, \sigma^w)$. Analogously denote $\pi^\tau(z,\sigma)$ the *expected utility* of a type τ agent when choosing action z in a population that plays σ. For a $0-$type material payoffs coincide with utility whereas for a $w-$type utility is in general distinct from material payoffs because of the psychological payoff loss. More precisely $\Pi^\tau(z,\sigma)$ and $\pi^\tau(z,\sigma)$ are given by the following expressions.

$$\begin{aligned}
\pi^0(C,\sigma) &= \Pi^0(C,\sigma) = a[(1-px)\sigma_C^0 + px\sigma_C^w] \\
\pi^0(D,\sigma) &= \Pi^0(D,\sigma) = (1-px)(\sigma_C^0 + \sigma_D^0 d) + px(\sigma_C^w + \sigma_D^w d) \\
\pi^w(C,\sigma) &= \Pi^w(C,\sigma) = a[(1-p)x\sigma_C^0 + (1-(1-p)x)\,\sigma_C^w] \\
\pi^w(D,\sigma) &= (1-p)x(\sigma_C^0(1-w) + \sigma_D^0(d-w)) \\
&\quad + (1-(1-p)x)\,(\sigma_C^w(1-w) + \sigma_D^w(d-w)) \\
\Pi^w(D,\sigma) &= (1-p)x(\sigma_C^0 + \sigma_D^0 d) + (1-(1-p)x)\,(\sigma_C^w + \sigma_D^w d)
\end{aligned}$$

Nash Equilibrium

Having specified payoffs we have a complete description of the population game $\Gamma = (I, T, Z, p, \pi(\cdot))$. To describe optimal behavior we rely on the concept of Nash-equilibrium.[10]

Definition 3 (Nash equilibrium) A Nash equilibrium of Γ is any population profile (σ^0, σ^w, p) such that $\sigma_z^\tau > 0 \Rightarrow z \in \arg\max_Z \pi^\tau(z,\sigma), \forall \tau \in T$.

We are now interested in how rational behavior in the population game affects the cultural transmission of social norms.

2.3 The cultural transmission process

Social norms are internalized via three mechanisms described below: vertical transmission (socialization through parents), horizontal transmission (peer interaction) and institutional transmission.[11]

[10] Again there is a formal equivalence between the Nash-equilibria of Γ and the symmetric Bayes-Nash equilibria of the bilateral game with incomplete information.

[11] The term oblique transmission is typically used instead of institutional transmission. We chose the latter because our focus is on intra-generational transmission. See Gintis (2003), Bisin et al (2004), Cavalli-Sforza and Feldmann (1981) or Cialdini and Trost (1998) among others.

Horizontal Transmission

Agents internalize social norms partly through social learning from peers. We assume that this process is payoff-biased (agents are more likely to adapt norms from materially successful agents).[12] As in indirect evolutionary models there is thus a distinction in our model between the payoffs that drive behavior and those that are relevant for selection.[13]

Definition 4 (Cultural Model) A cultural model of type m for agent i is another agent who reveals her preferences and material payoffs to i.

A cultural model can, for example, be a friend, a colleague or a partner. At any point in time t an individual of type τ meets a cultural model of type $m^t \in \{0, w\}$ according to the matching probabilities and observes the model's type and material payoff Π_t^m in that period.[14] Note that the cultural model is not the person one interacts with in the stage game. If an agent's cultural model is of his type he will stick to his norm with probability 1. If the cultural model is of another type he might adopt her norm with a probability that depends linearly on (positive) payoff-differences. The probability that an individual of type τ adapts the m- norm at time t is given by

$$\Pr(m^t|\tau)_t = \begin{cases} (1-\alpha) + \alpha(\Pi_t^m - \Pi_t^\tau)1_+ & \text{if } m^t \neq \tau \\ 1 & \text{if } m^t = \tau \end{cases} . \tag{3}$$

$\alpha \in (0,1)$ is a parameter that measures the importance of the payoff-bias in horizontal transmission. 1_+ is the indicator function taking the value 1 if the preceding term is positive and 0 otherwise.[15] The total share of w- types in the population after horizontal transmission is then given by

$$\begin{aligned} \Pr(w)_t &= p_t(1 - \Pr(0|w)_t) + (1-p_t)\Pr(w|0)_t \\ &= p_t + (1-p_t)p_t x\alpha(\Pi_t^w - \Pi_t^0). \end{aligned} \tag{4}$$

Accordingly

$$\Pr(0)_t = 1 - \Pr(w)_t \tag{5}$$

denotes the total share of 0- types after horizontal transmission.[16]

[12] There is evidence suggesting that agents are more likely to adopt norms from agents who are materially rich. See Henrich and Gil-White (2000) or Boyd and Richerson (2005).

[13] See Bester and Güth (1998) or Huck (1998) among others. Having selection work on utility (instead of material preferences) would make the modeling exercise meaningless, as the survival of a trait in this case depends on how strong it is *assumed* to be.

[14] We will omit the argument of the payoff function when it can be done without ambiguity.

[15] Such linear rules are standard in the literature on social learning. Properties of such rules are derived in Schlag (1998) or Manski (2004) among others.

[16] Some norms can be transmitted simply by observing other's behavior (see e.g. Cialdini et al (1990)). Note though that agents with different norms can display the same behavior in this model. That is why preferences have to be observed here for norm transmission. See Cialdini and Trost for a discussion. Many studies also show that transmission mainly operates between people who feel "close" to each other. Latané (1996) for example emphasizes the importance of communication for norm-transmission.

Institutional Transmission

The adoption of pro-social norms can be enhanced through the institutions of a society.[17] By structuring interactions institutions lead to framing and other situation construal effects that favor the spread of some social norms. Legal norms or government policies can stigmatize some behaviors while promoting others. The pro-social norm can also be transmitted through socialization institutions such as schools, universities or churches. Finally communication media can shift reference points and affect norm-transmission. Under the influence of institutional pressures some 0- types will switch to the w- norm. Institutional transmission is proportional to the "effective" number of w- types in the society (the number of w- types a 0- type perceives in his environment $p_t x$). The underlying idea is that institutional transmission is more effective if there are more w- types. Legal norms turn more easily into social norms, public policies are better implemented and norms more efficiently transmitted in schools if there are more w- types.[18] Having the parameter $\psi \in [0,1]$ measure the strength of institutional pressures the share of 0- types that adapt the w- norm because of institutional transmission is given by $p_t x \psi$.

Vertical Transmission

As the focus of the paper is on intra-generational, we assume that vertical transmission is unbiased in the sense that each $\tau-$type at his/her death leaves exactly one offspring of type τ.

Population Dynamics

Adding up we get the following population dynamics:

$$\begin{aligned} p_{t+1} &= \Pr(w)_t + p_t x\psi(1-\Pr(w)_t) \\ &= p_t + p_t(1-p_t)x[\alpha(\Pi_t^w - \Pi_t^0)(1-p_t x\psi)+\psi], \end{aligned}$$

or in the continuous time approximation,

$$\dot{p} = p(1-p)x[\alpha(\Pi^w - \Pi^0)(1-px\psi)+\psi] =: f(p). \tag{6}$$

Note that if $\psi = 0$ this equals the familiar replicator dynamics (up to a change of time scale). The cultural transmission process is illustrated in Figure 1.

Figure 1 about here

3 Cultural Equilibrium

We call a cultural equilibrium in this model a situation where - given equilibrium play in the population game - the share of $w-$ types in the population remains constant.

[17] Many studies show that ideals and norms are not absolute but influenced by the institutional structure in which an agent is placed. See Schotter et al (1996), Alesina and Fuchs-Schündeln (2005), Bowles (1998), Huck (1998) or Gintis (2003) among others.

[18] While it is clear that institutional transmission cannot be independent of the effective number of w- types, the assumption of exact proportionality is maybe the most conservative in an attempt to keep the number of parameters in the model limited.

Definition 5 (Cultural Equilibrium) *A cultural equilibrium is a population state p that satisfies $\dot{p} = 0$ in equation (6).*

Typically we will be interested in cultural equilibria that are locally asymptotically stable in the sense that the state trajectory always returns to the equilibrium state given that it is "close enough". The set of locally stable cultural equilibria obviously depends on the strength of the norm. We will describe the different cases in turn.

3.1 Weak Norm

If the norm is weak, defection is a dominant strategy in the bilateral game for both the w- and the 0- types. In this case the population dynamics is trivial; since payoffs for both types are the same, horizontal transmission is neutral and the dynamics is governed by institutional transmission only. Full internalization of the w- norm ($p = 1$) is globally stable whenever $(\psi, x) >> 0$. Note though that with weak norms norm - internalization leads to behavior that is "phenotypically" indistinguishable from behavior without the norm. Any cultural equilibrium will be characterized by full defection. As an illustration consider the example of charities in the US that make commercials using the slogan "almost giving is not enough". This could suggest that many people have internalized a weak norm for giving, believing that giving (C) is good (or not giving (D) bad), but not acting upon this belief.

3.2 Strict Norm

If the norm is strict, cooperation is a dominant strategy for the w- type. Consequently all the Nash-equilibria of the population game are of the form (D, C, p) i.e. population profiles where all 0- types play D and all w- types play C. The cultural equilibria in this case are both monomorphic states as well as the polymorphic states p_1 and p_2.[19] Which of these will be locally stable depends on the vector of institutional characteristics (ψ, x). It is clear that very high institutional pressures ψ always lead to the spread of the w- norm. Let us then focus first on the more interesting case where ψ is arbitrarily small.

If the degree of integration is very small (if $0 < x < \min\{\frac{a-d}{1-d}, 1 - \frac{d}{a}\}$) the monomorphic equilibrium $p = 1$ is globally stable, because for low x both types mainly interact among each other. As a consequence w- types will get the high payoff for joint cooperation relatively often while 0- types will often get the lower payoff associated with mutual defection. This leads to global convergence to $p = 1$.

If integration takes on intermediate values two mutually exclusive cases arise depending on the payoff parameters. Cooperation survives in both. If $x \in (\frac{a-d}{1-d}, 1-\frac{d}{a})$, the globally stable equilibrium is the polymorphic state p_1, since for low levels of norm - internalization, 0- types will obtain lower material payoffs

[19] The expressions for p_1 and p_2 are rather complicated and stated in Appendix A.

in expectation, which biases social learning in favor of the w- norm and has p rise. As p rises this payoff bias shrinks and reverts at p_1. If on the other hand $x \in (1-\frac{d}{a}, \frac{a-d}{1-d})$ this process goes the other way round. The polymorphic equilibrium will be unstable, and both monomorphic states will be locally stable with their basins of attraction separated by p_2.

Finally if the degree of integration is very high ($x > \max\{1-\frac{d}{a}, \frac{a-d}{1-d}\}$) 0-types will be able to benefit from the cooperative behavior of the w- types and thus obtain a higher material payoff.

Proposition 1 *If $w > \max\{1-a, d\}$, $\psi > 0$ arbitrarily small and*
(i) if $0 < x < \min\{\frac{a-d}{1-d}, 1-\frac{d}{a}\}$ the globally stable equilibrium is $p^* = 1$.
(ii) if $x \in (\frac{a-d}{1-d}, 1-\frac{d}{a})$ the globally stable equilibrium is $p^* = p_1$.
(iii)if $x \in (1-\frac{d}{a}, \frac{a-d}{1-d})$ the locally stable equilibria are $p^* = \{0, 1\}$.
(iv) if $x > \max\{1-\frac{d}{a}, \frac{a-d}{1-d}\}$ the globally stable equilibrium is $p^* = 0$.

Proof. See Appendix A. ■

The intuition for this result is clear. If the social norm is strict w - types cooperate unconditionally in the prisoners' dilemma. The norm (and thus cooperative behavior) survives if and only if the benefits of this altruistic behavior fall disproportionately onto other $w-$types. This is the case whenever the two types are sufficiently separated.

The more integrated societies are, the more institutional pressures are needed to sustain strict norms. In fact there are two critical levels of institutional pressures that can ensure the persistence of the pro-social norm. These threshold levels are given by $\psi_1 =: \frac{[x(1-d)-(a-d)]\alpha}{1-x[(a-d)-x(1-d)]\alpha}$ and $\psi_2 := \alpha[xa-(a-d)]$ for the two mutually exclusive parameter constellations where $a + d \lesseqgtr 1$. Note that both thresholds are strictly increasing with α and vanish if $\alpha = 0$. The reason is that for $\alpha = 0$ social learning displays no payoff-bias, but then any arbitrarily small level of institutional transmission will induce global convergence to $p = 1$. Note also that both thresholds rise with x. The intuition is that for strict norms more integration biases social learning against the norm. Consequently institutional pressures need to be higher to sustain it. Consider first the case where $a+d < 1$. This is the case where material gains of unilateral defection are higher than the opportunity costs of unilateral cooperation.

Corollary 1a *If $a+d < 1$ the monomorphic equilibrium state $p^* = 1$ is globally stable iff $\psi > \psi_1$.If $\psi \in [\psi_2, \psi_1]$ cooperation survives in the polymorphic equilibrium $p = p_1$.*

Proof. Appendix A ■

In the second case where $a + d > 1$ we have:

Corollary 1b *If $a+d > 1$ the monomorphic equilibrium state $p^* = 1$ is globally stable iff $\psi > \psi_2$. If $\psi \in [\psi_1, \psi_2]$ the monomorphism $p = 1$ is still locally stable.*

Proof. Appendix A ■

Figure 2 displays the state equation as a function of p and x for varying strengths of institutional pressures.[20] If ψ is small, as in Figure 2a, it is mainly the degree of integration of the society that acts as a selecting force to determine the set of locally stable equilibria of the system. It can be seen that for small x only $p^* = 1$ is locally stable, for intermediate x the globally stable equilibrium is polymorphic, and for high integration $p^* = 0$ is globally stable. In Figure 2b the forces of institutional pressures outweigh the forces of integration, so $p^* = 1$ is globally stable, but the speed of convergence is maximal for interior x. In Figure 2c institutional pressures dominate all other forces. Consequently $p^* = 1$ is selected, convergence being faster for higher levels of integration.

Summary *With strict norms cooperation either needs high separation or sufficiently strong institutional pressures to persist in a cultural population equilibrium.*

Figure 2a - 2c about here

As it should be clear by now that higher institutional pressures always enhance the evolutionary selection of the w- norm, we will focus in the following sections on the case where ψ is strictly positive but arbitrarily small.

3.3 Intermediate Norm

If the norm is intermediate in strength, two mutually exclusive cases can arise depending on the payoff-parameters.

3.3.1 $a + d > 1$: Prisoners and Stag-Hunters

In this case the norm is intermediate whenever $w \in [1 - a, d]$. The payoff table (2) then describes a stag-hunt game. Remember that this (bilateral) game has two Nash-equilibria in pure strategies in which players either both cooperate or both free-ride. To see what are the Nash equilibria of the population game first note that D is still a dominant strategy for a 0-type. Clearly then the profiles where both types defect ((D, D, p)) are Nash-equilibria $\forall p \in [0, 1]$. The profiles (D, C, p) where w- types play C (and 0-types D) will be equilibria if and only if

$$\pi^w(C, z^*) \geq \pi^w(D, z^*) \tag{7}$$

where $z^* = (z^0, z^w)^* = (D, C)$. This is equivalent to

$$p \geq \frac{(1 - w - a) - x(1 - d - a)}{x(a + d - 1)} =: \widetilde{p}. \tag{8}$$

We can state the following result.

Proposition 2 *If $w \in [1 - a, d]$ the Nash-equilibria of the population game Γ are given by $(D, D, p), \forall p \in [0, 1]$ and $(D, C, p), \forall p \in [\widetilde{p}, 1]$.*

[20] The parameters used for the graphs are: $a = 1/2$, $d = 1/4$, and $\alpha = 1/2$.

Obviously for $p \geq \widetilde{p}$ an issue of equilibrium selection arises. Two cases are possible.

Case 1a: If $p \geq \widetilde{p}$ the equilibrium selected is (D, D, p).

Case 1b: If $p \geq \widetilde{p}$ the equilibrium selected is (D, C, p).

In case 1a the parameter region here is indistinguishable from that of the weak norm and the analysis from sub-section 3.1 applies. Let us thus focus on case 1b.

If $x < \frac{a+w-1}{a+d-1}$, $\widetilde{p}$ is negative implying that a w- type will cooperate unconditionally for all population shares p. The analysis for this range of x thus corresponds to the case of the strict norm discussed above.

Focus on the case where $x \geq \frac{a+w-1}{a+d-1}$. Then there exists a non-empty range of population shares $[0, \widetilde{p})$ in which w- types will find it optimal to free-ride, in this way depriving the 0- types of their payoff advantage from unilateral defection. Consequently if $p < \widetilde{p}$ both types will earn the same material payoff in expectation and the dynamics of norm - internalization will be governed exclusively by institutional transmission. This leads to a steady growth in norm - internalization until the share of w- types reaches $\widetilde{p}$. Two cases arise: if $x \in (\frac{a+w-1}{a+d-1}, \frac{a-d}{1-d}]$ the globally stable equilibrium is $p = 1$, whereas if $x > \max\{\frac{a+w-1}{a+d-1}, \frac{a-d}{1-d}\}$ the payoff-bias works against the w- norm destabilizing $p = 1$ and stabilizing $p = \widetilde{p}$. Long-run cooperation is enhanced compared to the case of strict norms.

Proposition 3 *If* $w \in [1-a, d]$, $\psi > 0$ *arbitrarily small and*

(i) if $0 < x < \min\{\frac{a+w-1}{a+d-1}, \frac{a-d}{1-d}\}$ *the globally stable equilibrium is* $p^* = 1$.

(ii) if $x \in (\frac{a-d}{1-d}, \frac{a+w-1}{a+d-1})$ *the locally stable equilibria are* $p^* = \{0, 1\}$.

(iii) if $x \in [\frac{a+w-1}{a+d-1}, \frac{a-d}{1-d}]$ *the globally stable equilibrium is* $p^* = 1$.

(ii) if $x > \max\{\frac{a+w-1}{a+d-1}, \frac{a-d}{1-d}\}$ *the globally stable equilibrium is* $p^* = \widetilde{p}$.

Proof. *Appendix A* ■

Under the conditions of this proposition and if the degree of integration is sufficiently high ($x > \frac{a+w-1}{a+d-1}$), every stable equilibrium involves cooperation (even if $(\psi, x) \to (0, 1)$). This is in stark contrast to the case of the strict norm. With strict norms as $(\psi, x) \to (0, 1)$ the set of locally stable equilibria reduces to $p^* = 0$. Norms of intermediate strength though will always survive in fully integrated societies, because now w- types are conditional cooperators, cooperating only if the share of w- types p is high enough. Consequently they cannot be as easily exploited by non-cooperators. Note also that $\widetilde{p}$ rises with the degree of integration x. There will be more cooperation in any stable polymorphic cultural equilibrium for higher degrees of integration.

3.3.2 $a + d < 1$: Prisoners and Chickens

The intermediate norm corresponding to this case is $w \in [d, 1-a]$. The game form (2) then represents a chicken game. This (bilateral) game has two asymmetric Nash-equilibria in pure strategies in which one player plays C and one player D. This has as a consequence that in a population with "many" w-types, there is no Nash-equilibrium where all w- types choose the same action

z. In any cultural equilibrium in this region a w- type will randomize. If on the other hand the share of 0-types is sufficiently high, a w- type will find it optimal to play C. (As in this case he is matched with high probability with a 0- type who has as a dominant strategy to play D.) This case occurs whenever

$$\pi^w(C, z^*) \geq \pi^w(D, z^*) \tag{9}$$

where $z^* = (D, C)$ or equivalently iff

$$p \leq \frac{(1-w-a) - x(1-d-a)}{x(a+d-1)} = \widetilde{p}.$$

We can state the following result.

Proposition 4 *If $w \in [d, 1-a]$ the Nash-equilibria of the population game Γ are given by: $(D, C, p), \forall p \in [0, \widetilde{p}]$ and $(D, (\sigma_C^{w*}, (1-\sigma_C^{w*}), p) \forall p \in (\widetilde{p}, 1]$, where $\sigma_C^{w*} = \frac{w-d}{(1-(1-p)x)[1-a-d]}$.*

Proof. Appendix A ■

Now w- types cooperate (play C) if there is a low level of norm - internalization and randomize if p is high. Again for high degrees of separation full internalization of the w- norm ($p = 1$) is globally stable since in this case w-types will be mainly matched with other w- types. For intermediate degrees of integration, $a + d < 1$ (meaning that the gain of unilateral defection is higher than the (opportunity) cost of unilateral cooperation) implies that if w- types are matched mainly with each other and if $p \geq \widetilde{p}$ such that they use the mixed action $\sigma^{w*} = (\sigma_C^{w*}, (1-\sigma_C^{w*}))$, they will obtain a higher payoff on average than 0- types mainly matched with each other. As the degree of integration rises this material payoff advantage will diminish and finally reverse in favor of the 0- types. For $p < \widetilde{p}$ w- types will cooperate and obtain lower material payoffs than 0- types whenever integration is high. Consequently for very high degrees of integration $p = 0$ is globally stable.

Proposition 5 *If $w \in [d, 1-a]$,$\psi > 0$ arbitrarily small and*
(i) if $0 < x < \min\{1 - \frac{d}{a}, \frac{1-d-w}{1-d}\}$ the globally stable equilibrium is $p^* = 1$.
(ii) if $x \in [1 - \frac{d}{a}, \frac{1-d-w}{1-d})$ the locally stable equilibria are $p^* = \{0, 1\}$.
(iii) if $x \in [\frac{1-d-w}{1-d}, 1 - \frac{d}{a}]$ the globally stable equilibrium is $p^* = p_1$.
(iv) if $x > \max\{\frac{1-d-w}{1-d}, 1 - \frac{d}{a}\}$ the globally stable equilibrium is $p^* = 0$.
Proof. *Appendix A* ■

Now the pro-social norm does not survive as a preference trait in fully integrated societies (as $(\psi, x) \to (0, 1)$), even though w- types are conditional cooperators. The reason lies in the fact that now w- types find it optimal to cooperate whenever they are few. This perhaps somewhat paradoxical result comes from the incentives the payoffs in the chicken game provide. Let us compare these incentives to those in the stag-hunt game: In the stag-hunt game establishing joint cooperation is difficult because of "fear". A w- type fears that

whenever he plays C he could be matched with someone playing D and in this way be exploited. On the contrary in the chicken game the problem is "greed" rather than "fear"; a w- type matched with someone who cooperates wants to play D because unilateral defection is still profitable in spite of the existence of the pro-social norm. In the stag hunt game higher shares of norm - internalization enhance cooperation by w- types, because a high share of w- types can reduce the fear of being exploited (making this more unlikely). In the chicken game context it is a high population share of 0- types that enhances cooperation by w- types because the probability of the match defecting is high. This renders $p = 1$ unstable in integrated societies while making $p = 0$ a global attractor. Note also that in the case of the chicken game full norm internalization ($p = 1$) does not mean that everyone will cooperate in equilibrium. If $p = 1$ the level of cooperation in the population will be $\sigma^* = \frac{w-d}{1-a-d}$. [21]

Summary *If norms are of intermediate strength cooperation can survive in both scenarios: high separation and high integration. High institutional pressures are necessary for the survival of cooperation in integrated societies under some parameter constellations but not under others.*

4 Endogenous Norm-strength

The baseline case of exogenous norms illustrates that norm strength matters when it comes to determining the equilibrium share of w- types. For some norms though, norm strength will not be exogenous. Rather it will depend on the informational environment, such as, for example, the distribution of preferences in an agent's sample. In this section we endogenize norm strength by linking it to the share of w- types in society.[22]

In particular we will assume that the strength of internal punishment rises with the number of w- types in the sample of a particular w- type. "It's not right what I'm doing, but as nobody else cares, it's ok" is a revealing phrase that often accompanies norm-guided behavior. Well-known examples where the fact that norm-internalization is low reduces the strength of internal sanctions include not going to vote, minor tax evasion, welfare dependency, not going to church, divorce or free-riding on public transport. Consider the example of a w- type thinking that tax evasion is "bad". If he is surrounded by 0- types who argue that tax evasion is a rational reaction to a badly designed system and thus "not bad", his norm will be weakened and the psychological payoff loss upon evading taxes smaller. If on the other hand he is surrounded by w- types who argue (like him) that tax evasion is stealing and thus "bad", his norm will be strengthened and the psychological payoff loss higher.[23]

[21] The level of cooperation as a function of p is given in Proposition 4.

[22] Of course one could also want to endogenize x instead of or in addition to w. This would lead to a non-trivial optimal control problem for a social planner. In addition it is not clear what the objective function of such a planner should be, as there are several problems with welfare measurement in this context (see an earlier version of this paper (Mengel (2006))).

[23] Note that in our society all w- types face the same distribution of types allowing us to

To formalize this idea denote the proportion of w- types in a w- type's sample by $s := [1-(1-p)x]$ and let the strength of the norm be given by some function

$$w(s) : [0,1] \to [0,1]$$

such that $w(1) = 1$, $w(0) = 0$, $w(s) \in C^2$ and $\frac{\partial w(s)}{\partial s} > 0$. The sign of the derivative expresses the fact that more norm - internalization tends to make a norm stronger. The cultural equilibrium determines the strength of the norm. On the other hand the strength of a social norm affects peoples' preferences, actions and the likelihood that the norm is internalized. In this way the strength of the social norm determines the cultural equilibrium. This sort of feedback-effects between equilibrium and social norm are in many cases characteristic for norm-guided behavior. By focusing on only one of the two aspects, equilibrium or norm, one can miss an important part of the picture.

The change in norm strength is linked to the evolution of norm - internalization as follows:

$$\dot{w} = \frac{\partial w(s)}{\partial s} x \dot{p} \tag{10}$$

It can be seen that separation increases norm strength($\frac{\partial s}{\partial x} < 0$). Higher separation implies that w- types mainly interact among each other, so in each $w-$ type's sample the share of w- types will be very high and consequently the norm very strict. This fact will strongly impact our previous results.

Consider first the case where the payoff matrix (1) is such that $a + d > 1$. If the degree of integration is low, w- types almost exclusively interact with other w- types. This implies that the share of w- types in any w- type's sample is high, the social norm strict and thus (as we know from Section 3.2) only sustainable through very high degrees of separation. In this sense rigidity is self-reinforcing. Rigidity (separation) leads to strict norms, which in turn need even more rigidity to persist.

Whenever $0 < x < \min\{1 - \frac{d}{a}, \frac{a+w(\tilde{s})-1}{a+d-1}\}$ the society is sufficiently separated to sustain strict norms, as the benefits of pro-social behavior fall disproportionately on w- types.[24] In this parameter range the globally stable equilibrium is $p^* = 1$.

Slightly higher degrees of integration will still lead to strict norms, but not anymore to a material payoff-advantage for w- types. Consequently the norm will not be selected by the evolutionary dynamics. As the degree of integration further rises, norm strength will fall. Finally high degrees of integration will lead to intermediate norms sustained in polymorphic equilibria.

Note that if x is very high both monomorphic equilibria are unstable. The reason is that if $p \to 1$ the norm will be strict and thus not sustainable with high integration. On the other hand if $p \to 0$ the norm becomes weak, driving the dynamics away from $p = 0$. Fully integrated societies thus sustain a globally stable polymorphic equilibrium with intermediate norm strength.

continue to treat the norm as a unique variable.

[24] $\tilde{s}$ denotes the solution to the following fixed point equation: $\frac{(1-w(s)-a)-x(1-d-a)}{x(a+d-1)} = p$.

Proposition 6 *If* $a+d>1, \psi \to 0$ *and*

(i) $0<x<\min\{1-\frac{d}{a}, \frac{a+w(\widetilde{s})-1}{a+d-1}\}$, *the globally stable equilibrium is* $p^*=1$.

(ii) $1-\frac{d}{a}<x<\min\{\frac{a+w(\widetilde{s})-1}{a+d-1}, \frac{a-d}{1-d}\}$, *the stable equilibria are* $p^*=\{0,1\}$.

(iii) $\frac{a-d}{1-d}<x<\frac{a+w(\widetilde{s})-1}{a+d-1}$, *the globally stable equilibrium is* $p^*=0$.

(iv) $\frac{a+w(\widetilde{s})-1}{a+d-1}<x<\frac{a-d}{1-d}$, *the locally stable equilibria are* $p^*=\{\widetilde{p},1\}$.

(v) $x>\max\{\frac{a-d}{1-d}, \frac{a+w(\widetilde{s})-1}{a+d-1}\}$, *the globally stable equilibrium is* $p^*=\widetilde{p}$.

Proof. Appendix B ■

There are two scenarios in which cooperation survives in a globally stable equilibrium: with high separation sustained by strict norms and corresponding high levels of internal punishment, and in very integrated societies sustained by intermediate norms and correspondingly lower levels of internal punishment. Note that only the latter equilibria are polymorphic. Maybe somewhat counter-intuitively, integrated societies sustain heterogeneity while separated societies imply monomorphic equilibria. Also note that the share of w- types in the polymorphic equilibrium is maximized at $x=1$.

With endogenous norm strengths the relation between integration and norm-internalization (and thus cooperation) is not monotone. The reason is that integration affects preferences via two channels. It affects behavior and thus norm-internalization, but this has feedback effects on the strength of the social norm itself. If separation is high these feedback effects can be so strong that agents having internalized the norm lose their ability to react to exploitation (cooperating will be a dominant strategy for them). In these cases even more separation is needed to protect the norm. The rigid population structure characterized by high separation leads to strict norms, which need even more rigidity to survive. Rigidity is self-reinforcing. The following example illustrates these results.

Example I *Consider the case where norm strength depends linearly on norm-internalization (i.e. where* $w(s)=s$*). Assume that* $a=3/4$ *and* $d=1/2$*. In this case* $a+d>1$ *(for a* w*- type the loss of unilateral cooperation is higher than the gain of unilateral defection). As can be seen in Figure 3 for* $x<1/3$*, the norm is strict in equilibrium (*$w=1$*) and* $p=1$ *is globally stable. For* $x\in[1/3,1/2]$ *both monomorphic equilibria are locally stable with strict norms in both cases. In the equilibrium* $p=0$ *norm strength is linearly decreasing in* x *(*$w=1-x$*). For* $x\in(1/2,3/5]$ *norm strength is intermediate but only* $p=0$ *is locally stable. The reason is that for* $x<3/5$ w- types *are unconditional cooperators even for intermediate norm-strengths. Finally for* $x>3/5$ *norm strength is intermediate* ($w=2/5$) *and the polymorphic equilibrium* $\widetilde{p}=1-\frac{3}{5x}$ *is globally stable.*

Figure 3a - 3b about here

The case in which $a+d<1$ (where the material loss of unilateral cooperation is smaller than the gain of unilateral defection) delivers qualitatively the same

result.[25] Cooperation survives in a very separated society sustained by strict norms in a monomorphic equilibrium and in very integrated societies sustained by intermediate norms in a polymorphic equilibrium given by $\widehat{p} := 1 - \frac{1-w^{-1}(d)}{x}$.

Proposition 7 *If* $a + d < 1, \psi \to 0$ *and*
(i) $x < \frac{a-d}{1-d}$, *the globally stable equilibrium is* $p^* = 1$.
(ii) $x \in [\frac{a-d}{1-d}, \min\{1 - \frac{d}{a}, 1 - w^{-1}(d)\}]$, *the glob. stable equilibrium is* $p^* = p_1$.
(iii) $x \in [1 - \frac{d}{a}, 1 - w^{-1}(d)]$, *the globally stable equilibrium is* $p^* = 0$.
(iv) $x > \max\{1 - w^{-1}(d), \frac{a-d}{1-d}\}$, *the globally stable equilibrium is* $p^* = \widehat{p}$.

Proof. Appendix B ■

Note that for both cases $a + d \lesseqgtr 1$ - contrary to what is obtained with standard direct evolutionary mechanisms - the long-run equilibrium in fully integrated societies (where matching is random) is always polymorphic. Furthermore in all these polymorphic equilibria w- types are conditional cooperators, and a positive level of overall cooperation is observed.[26] This is a behavioral pattern that is found also in many experimental studies on cooperation problems in western societies.[27]

Summary *With endogenous norm strength and for vanishingly low levels of institutional pressures, cooperation always survives in two scenarios, with high separation sustained by strict norms in monomorphic equilibria, and in very integrated societies sustained by intermediate norms in polymorphic equilibria.*

5 Conclusions

In this paper we propose and study a cultural selection mechanism for preference traits. In particular we concentrate on social norms for cooperation and ask under which conditions pro-social norms can survive if not all agents have internalized these norms. The main question examined is how the institutions of a society and in particular the degree of integration impact norm internalization in the long run.

We find that strict norms for cooperation need either separation or strong institutional pressures in order to survive. On the contrary intermediate norms can survive even in completely integrated societies and with vanishingly low levels of institutional pressures. Endogenizing the strength of the norm we find that there are two scenarios under which cooperation can survive. The first scenario is that of a rigid society, displaying a high degree of separation and very strict

[25] The case $a + d = 1$ is described at the end of Appendix B.

[26] In the case of Proposition 6 $\sigma^* = \widetilde{p}$, and in the case of Proposition 7 σ^* can be obtained from the equation given in Propositon 4 (substituting $\widehat{p}$).

[27] Fischbacher et al (2001) find that roughly 50% of the participants in their public goods experiment are conditional cooperators, 30% always free-ride and only very few cooperate unconditionally. See also Grimm and Mengel (2007a, 2007b) and the references contained therein.

norms sustained by strong internal punishment. Cooperation in this scenario is achieved through rigid population structures that in turn lead to strict norms. The second scenario is one of an integrated society with intermediate norms sustained by lower internal punishment and displaying heterogeneity of types in equilibrium. Here integration stabilizes a polymorphic equilibrium with norms that are not as strict. In fact in fully integrated societies all stable equilibria are polymorphic and there is conditional cooperation. This contrasts with results obtained by relying on standard direct evolutionary mechanisms but is in line with experimental results.

Our findings show that the relation between integration and cooperation is not as simple as commonly assumed in the literature that centers around the group-selection idea.[28] Whether higher separation (locally) helps or hurts cooperation depends on how strict norms are. In particular if norm strength is endogenous the relation between integration and cooperation is non-monotone and more separation will often be detrimental to cooperation. The reason is that if cooperation is sustained through social norms there can be feedback effects from the interaction structure on the norm itself. High separation can render norms so strict that agents having internalized these norms lose their ability to react to exploitation. Given the recent revival of group selection ideas, it is important to delimit the context in which these results obtain carefully.

References

Alesina, A., Fuchs-Schündeln, N. 2005. Good bye Lenin (or not ?) - The effect of communism on people's preferences. NBER working paper 11700.

Azar, O. H. 2001. What sustains social norms and how they evolve ? Journal of Economic Behavior and Organization 54, 49-64.

Benabou, R., Tirole, J. 2006. Incentives and prosocial behavior. American Economic Review 96, 1652-1678.

Bester, H., Güth, W. 1998. Is altruism evolutionary stable. Journal of Economic Behavior and Organization 34, 193-209.

Bisin, A., Topa, G., Verdier, T. 2004, Cooperation as a transmitted cultural trait, Rationality and Society 16, 477-507.

Boyd, R., Richerson, P. 1990. Group selection among alternative evolutionary stable strategies. Journal of Theoretical Biology 145, 331-342.

Boyd, R., Richerson, P. 2005. The Origin and Evolution of Cultures (Evolution and Cognition). Chicago: University of Chicago Press.

[28] See for example Bowles and Gintis (1998), Boyd and Richerson (2005), Henrich (2004) or Mitteldorf and Wilson (2000) among many others.

Bowles, S. 1998. Endogenous preferences: The cultural consequences of markets and other economic institutions. Journal of Economic Literature 36, 75-111.

Bowles, S., Gintis, H. 1998. The moral economy of communities: structured populations and the evolution of pro-social norms. Evolution and Human Behavior 19, 3-25.

Cavalli-Sforza, L., Feldman, M. 1981. Cultural Transmission and Evolution. Princeton: Princeton University Press.

Cialdini, R., Raymond, B., Reno, R., Kallgreen, C., 1990. A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places. Journal of Personality and Social Psychology 58, 1015-1026.

Cialdini, R., Trost, M. 1998, Social influence, social norms, conformity and compliance. In: Gilbert, D.T., Fiske, S.T., Lindsey, G. (eds.), The Handbook of Social Psychology vol.2, Boston: Mc Graw-Hill, 151-192.

Elster, J. 1989. Social norms and economic theory. Journal of Economic Perspectives 3, 99-117.

Fischbacher, U., Gächter, S., Fehr, E. 2001. Are people conditionally cooperative ? Evidence from a public goods experiment. Economics Letters 71, 397-404.

Gintis, H. 2003. Solving the puzzle of pro-sociality, Rationality and Society 152, 155-187.

Grasmick, H., Green, D. 1980. Legal punishment, social disapproval and internalization as inhibitors of illegal behaviour. Journal of Criminal Law and Criminology 71, 325-335.

Grimm, V., Mengel, F. 2007a. Group selection with imperfect separation - an experiment, IVIE working paper AD 2007-05.

Grimm, V., Mengel, F. 2007b. Cooperation in viscous populations - experimental evidence, IVIE-working paper AD 2007-17.

Guttman, J. 2001a. Self-enforcing reciprocity norms and intergenerational transfers: theory and evidence. Journal of Public Economics, 81,117-151.

Guttman, J. 2001b. Families, markets and self-enforcing reciprocity norms. Annales d'Economie et de Statistique 63/64, 89-110.

Guttman, J. 2003. Repeated interaction and the evolution of preferences for reciprocity. Economic Journal 113, 631-656.

Hamilton, W.D. 1964. The genetical evolution of social behaviour. Journal of Theoretical Biology 7, 1-52.

Henrich, J. ,Gil-White, F. 2000. The evolution of prestige. Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. Evolution and Human Behavior 22, 165-196.

Henrich, J. Boyd, R. 2001. Why people punish defectors. Journal of Theoretical Biology 208, 79-89.

Henrich, J. 2004. Cultural group selection, coevolutionary processes and large-scale cooperation. Journal of Economic Behavior and Organization 53, 3-35 and 127-143.

Huck, S. 1998. Trust, treason and trials: an example of how the evolution of preferences can be driven by legal institutions. Journal of Law, Economics and Organization, V1411.

Latané, B. 1996. Dynamic social impact - the creation of culture by communication, Journal of Communication 46, 13-25.

Lindbeck, A., Nyberg, S., Weibull, J. 1999. Social norms and economic incentives in the welfare state. Quarterly Journal of Economics 114, 1-35.

Liu, R. X. 2003. The moderating effect of internal and perceived external sanction threats on the relationship between deviant peer associations and criminal offending. Western Criminology Review 43, 191-202.

Manski, C. 2004. Social learning from private experience: the dynamics of the selection problem. Review of Economic Studies 71, 443-458.

Mengel, F. 2006. A model of immigration, integration and cultural transmission of social norms. IVIE working paper AD 2006-08.

Mitteldorf, J., Wilson, D.S. 2000. Population viscosity and the evolution of altruism. Journal of Theoretical Biology 204, 481-496.

Myerson R. B., Pollock, G.B., Swinkels, J.M. 1991. Viscous population equilibria. Games and Economic Behavior 3, 101-109.

Nyborg, K., Rege, M. 2003. On social norms: the evolution of considerate smoking behaviour. Journal of Economic Behavior and Organization 52, 323-340.

Price, G. 1970. Selection and covariance. Nature 227, 520-521.

Reno, R., Robert, R., Cialdini, B., Kallgreen, C. 1993. The transsituational influence of social norms. Journal of Personality and Social Psychology 64, 104-112.

Schlag, K. 1998. Why imitate and if so how ? A boundedly rational approach to multi-armed bandits. Journal of Economic Theory 78, 130-156.

Schotter, A., Weiss, A., Zapater, I. 1996. Fairness and survival in ultimatum dictatorship game. Journal of Economic Behavior and Organization 31, 37-56.

Traxler, C. 2005. Social norms, voting and the provision of public goods. mimeo University of Munich.

Vega-Redondo, F. 1996. Evolution, Games and Economic Behaviour. Oxford: Oxford University Press.

Weibull, J. 1995. Evolutionary Game Theory. Cambridge: MIT-Press.

Wilson, D.S., Sober, E. 1994. Re-introducing group selection to the human behavioural sciences. Behavioral and Brain Science 17, 585-654.

Young, P. 1998. Social norms and economic welfare. European Economic Review 42, 821-830.

A Appendix A (Exogenous Norm strength)

Proof of Proposition 1:
Proof. Assume $a+d \neq 1$. (The case $a+d=1$ is treated below). There are four zeros of (6): $p^*=0, p^*=1$ and

$$\begin{aligned} p^*_{1/2} &= \frac{(1-a-d)+\psi(a(1-x)-d)}{2(\psi x(1-a-d))} \\ &\mp \frac{\sqrt{\begin{array}{c}[(a+d-1)\alpha x-\alpha x\psi(a(1-x)-d)]^2 \\ -4(\alpha(a(1-x)-d)+\psi)(\alpha\psi x^2(1-a-d)\end{array}}}{2(\alpha\psi x^2(1-a-d))}. \end{aligned}$$

The derivative of the state equation evaluated at the two monomorphic equilibria is given by

$$f'(p)_{|p=0} = x(\psi+\alpha((1-x)a-d)$$

$$f'(p)_{|p=1} = -x[\psi+\alpha(1-\psi x)((a-d)-x(1-d))].^{29}$$

$f'(p)_{|p_{1/2}}$ is a complicated expression, but we know that if $a+d \lesseqgtr 1$,

$$\lim_{\psi\to 0} p_{1/2} = \frac{a(1-x)-d}{(1-a-d)x} =: p_0$$

whereas the other zero diverges ($\lim_{\psi\to 0} p_{2/1}=\infty$)

Furthermore we have that $f(p,\psi)$ as given by (6) converges uniformly to $f(p,0)=p(1-p)x\alpha(\Pi^w-\Pi^0)$ as $\psi\to 0$. This can be seen by noting that

$$\begin{aligned} & |f(p,\psi)-f(p,0)| \\ = \;& p(1-p)x\psi(1-px(\Pi^w-\Pi^0)) \\ \leq \;& \frac{\psi}{4}, \forall p\in[0,1]. \end{aligned}$$

In addition $f'(p,\psi) \overset{uniformly}{\rightarrow} f'(p,0)$. This allows us to write

$$\begin{aligned}\lim_{\psi\to 0} f'(p,\psi)_{|p_{1/2}} &= f'(p,0)_{|\lim_{\psi\to 0} p_{1/2}=p_0} \\ &= -\alpha\frac{[(a-d)-xa][x(1-d)-(a-d)]}{1-d-a}.\end{aligned}$$

Then it is easy to see that $p^*=0$ is locally stable iff

$$0<\psi<\alpha[d-(1-x)a] := \psi_2. \tag{11}$$

For $\psi\to 0$ this condition reduces to $x>1-d/a$ and $p^*=1$ is locally stable iff

$$((a-d)-x(1-d)>0)\vee(\psi>\frac{\alpha(x(1-d)-(a-d))}{1-\alpha x((a-d)-x(1-d))} =: \psi_1). \tag{12}$$

Again for $\psi\to 0$ this condition reduces to $x<(a-d)/(1-d)$ and $p_{1/2}$ is locally stable iff

$$-\alpha\frac{[(a-d)-xa][x(1-d)-(a-d)]}{1-d-a}<0. \tag{13}$$

Let us consider the four cases of Proposition 1. (i) In this parameter range $f'(p)_{|p=0}>0$ and $f'(p)_{|p=1}<0$, so we have that $p^*=0$ is unstable and $p^*=1$ is locally stable. Continuity of $f(p)$ implies that the number of regular interior equilibria has to be even. As $\alpha(\Pi^w-\Pi^0)(1-p_t x\Delta)+\psi =: \Phi(p,\psi)$ is a quadratic polynomial in p for any given ψ there are at most two regular interior equilibria. Two constellations of the payoff parameters have to be distinguished: if $a+d<1$ we have that $p_2>1$ and if $a+d>1\Rightarrow p_1<0$. As there can neither be exactly two nor exactly one interior solution, there has to be none. (ii) For the second part observe that in this parameter range $f'(p)_{|p=0}>0$,$\forall\psi\in[0,1]$ while $f'(p)_{|p=1}<0$ iff $\psi\geq\psi_1$. For ψ arbitrarily small both monomorphic equilibria are thus unstable. $p_1\in(0,1)$ and p_2 diverges as $a+d<1$. Also note that the number of interior equilibria has to be odd. (iii) Observe that in this parameter range $f'(p)_{|p=1}<0$ whereas $f'(p)_{|p=0}<0$ iff $\psi\leq\psi_2$. For $\psi\to 0$, $p^*=1$ and $p^*=0$ are stable. The interior equilibrium p_2 is unstable (as $a+d>1$) and separates the basins of attraction of the two locally stable equilibria. (iv) In this region $f'(p)_{|p=1}<0$ whenever $\psi\geq\psi_2$, while $f'(p)_{|p=0}<0$ iff $\psi\leq\psi_1$. For arbitrarily small ψ it is clear that only $p=0$ is stable. Interior equilibria are unstable. ■

Statement of result Case $\mathbf{a+d=1}$:

Whenever $x<1-\frac{d}{a}(=\frac{a-d}{1-d})$ the unique stable equilibrium is $p^*=1$ and whenever $x>1-\frac{d}{a}$ the unique stable equilibrium is given by

$$p^*=\begin{cases}1 & \text{if } \psi\geq\psi_1\\ 0 & \text{if } \psi<\psi_1\end{cases}.$$

Proof of Corollary 1a:

Proof. "If": It follows from (12) that $\psi > \psi_1$ is sufficient for local stability of $p = 1$. $\psi \in [\psi_2, \psi_1]$ implies that both monomorphic states are unstable. Exactly one regular interior zero thus exists. We know that if $a+d < 1$ this polymorphic equilibrium is locally stable. "Only if": Local stability of $p = 1$ implies either $x < \frac{a-d}{1-d}$ or $\psi > \psi_1$. But $x < \frac{a-d}{1-d}$ implies $\psi_1 < 0$. ■

Proof of Corollary 1b:

Proof. "If": It follows from (12) that $\psi > \psi_2 > \psi_1$ is sufficient for local stability of $p = 1$. $\psi \in [\psi_1, \psi_2]$ implies that both monomorphic states are locally stable. "Only if": Global stability of $p = 1$ is sufficient for $x < 1 - \frac{d}{a} < \frac{a-d}{1-d}$, but $x < 1 - \frac{d}{a}$ implies $\psi_2 < 0$. ■

Proof of Proposition 3:

Proof. First note that (D, C, p) is a Nash-equilibrium iff $\pi_t^w(C, z^*) \geq \pi_t^w(D, z^*)$ where $z^* = (D, C)$ or equivalently if and only if

$$\begin{aligned} [1-(1-p)x]a &\geq [1-(1-p)x](1-w)+(1-p)x(d-w) \\ &\Leftrightarrow \quad p \geq \frac{(1-w-a)-x(1-d-a)}{x(a+d-1)} =: \widetilde{p} \leq 1. \end{aligned}$$

$\widetilde{p} > 0$ iff $x > \frac{a+w-1}{a+d-1} \in [0, 1]$. In case 1b the population dynamics is then given by

$$\dot{p} = \begin{cases} p(1-p)x\Delta & \text{if } p < \widetilde{p} \\ p(1-p)x[\alpha(\Pi^w - \Pi^0)(1-px\Delta)+\psi] & \text{if } p \geq \widetilde{p} \end{cases}.$$

In the case of arbitrarily small ψ there are two zeros of this dynamics: $p^* = 0$ and $p^* = 1$. Note that $\lim_{\psi \to 0} p_2 = p_0 < \widetilde{p}$ and $\lim_{\psi \to 0} p_1 = \infty$. The derivative of the state equation is

$$f'(p) = \begin{cases} (1-2p)x\Delta & \text{if } p < \widetilde{p} \\ (1-2p)x[\alpha(\Pi^w - \Pi^0)(1-px\Delta)+\psi] & \\ \quad +p(1-p)x(-x\Delta\alpha(\Pi^w - \Pi^0)) & \text{if } p \geq \widetilde{p} \end{cases}.$$

Note that $p = 0$ is unstable whenever $\widetilde{p} > 0$ and $x > 0$ as in this case $f'(p)_{|p=0} = x\Delta > 0$. Furthermore we know that $p = 1$ is locally stable iff

$$\begin{aligned} f'(p)_{|p=1} &= -x[\psi + \alpha(1-x\Delta)(a-d(1-x)-x)] < 0 \\ &\Longleftrightarrow \quad x < \frac{a-d}{1-d} \vee [x > \frac{a-d}{1-d} \wedge \psi > \psi_1]. \end{aligned}$$

Remember that $x \leq \frac{a+w-1}{a+d-1} \Leftrightarrow \widetilde{p} < 0$. If this is the case w-types are unconditional cooperators and the proof of case (i) and case (ii) can be read directly from the Proof of Proposition 1.

Case (iii): $x \in (\frac{a+w-1}{a+d-1}, \frac{a-d}{1-d}]$. We have that $x > \frac{a+w-1}{a+d-1} \Rightarrow \widetilde{p} > 0 \Rightarrow f'(p)_{|p=0} > 0$ and $x \leq \frac{a-d}{1-d} \Rightarrow f'(p)_{|p=1} < 0$. Consequently $p = 0$ is unstable ($\dot{p} > 0 \; \forall p < \widetilde{p}$) and $p = 1$ globally stable.

Case (iv): $x > \max\{\frac{a+w-1}{a+d-1}, \frac{a-d}{1-d}\}$. We have that $x > \frac{a+w-1}{a+d-1} \Rightarrow \widetilde{p} > 0 \Rightarrow f'(p)_{|p=0} > 0$ and $x > \frac{a-d}{1-d} \Rightarrow f'(p)_{|p=1} < 0$. Consequently both $p = 0$ and $p = 1$ are unstable. As furthermore there is no interior regular equilibrium, $\widetilde{p}$ is stable with basin of attraction $[0, 1]$. ■

Proof of Proposition 4 and 5:

Proof. First note that (D, C, p) is a Nash-equilibrium iff $\pi_t^w(C, z^*) \geq \pi_t^w(D, z^*)$ where $z^* = (D, C)$

$$\begin{aligned} &\Leftrightarrow \quad [1-(1-p)x]a \geq [1-(1-p)x](1-w) + (1-p)x(d-w) \\ &\Leftrightarrow \quad p \leq \frac{(1-w-a) - x(1-d-a)}{x(a+d-1)} =: \widetilde{p} \leq 1. \end{aligned}$$

If $p \geq \widetilde{p}$ w-types will randomize using action $\sigma_w^* = (\sigma_C^{w*}, (1-\sigma_C^{w*}))$. $\pi_t^w(C, \sigma) = \pi_t^w(D, \sigma)$ implies

$$\begin{aligned} [1-(1-p)x]\sigma_C^w a \quad &= \quad [1-(1-p)x][\sigma_C^w(1-w) + (1-\sigma_C^w)(d-w)] \\ &\quad\quad +(1-p)x(d-w) \\ &\Longleftrightarrow \quad \sigma_C^{w*} = \frac{w-d}{[1-(1-p)x](1-a-d)}. \end{aligned}$$

Exspected material payoffs of a w-type are thus given by

$$\Pi^w = \begin{cases} [1-(1-p_t)x]a & \text{if } p_t \leq \widetilde{p} \\ \frac{ad-w(1-w)-(1-p)(ad-(1-d)w)x}{(a+d-1)(1-(1-p)x)} & \text{if } p_t > \widetilde{p} \end{cases}. \tag{14}$$

The exspected material payoff of a 0-type is

$$\Pi^0 = \begin{cases} p_t x + (1-p_t x)d & \text{if } p_t \leq \widetilde{p} \\ d + \frac{(1-d)(w-d)px}{(1-a-d)(1-(1-p)x)} & \text{if } p_t > \widetilde{p} \end{cases}. \tag{15}$$

Inserting into the population dynamics gives

$$f'(p, \psi)_{|p=0} = x(\psi + \alpha((1-x)a - d)$$

and

$$f'(p, \psi)_{|p=1} = -\frac{\alpha(w-d)x(1-d-w-x(1-d))}{1-a-d}.$$

Then $p = 0$ is locally stable iff $0 < \psi < \alpha(d-(1-x)a) = \psi_1$. $p = 1$ is locally stable iff $x < \frac{1-d-w}{1-d}$.It can be easily seen that in the limit where $\psi \to 0$ no interior regular equilibrium exists for the region where $p \geq \widetilde{p}$. In the region where $p < \widetilde{p}$ the unique regular interior equilibrium is given by p_1. Remember that $\lim_{\psi \to 0} p_1 = p_0 < \widetilde{p}, \lim_{\psi \to 0} p_2 = \infty$ and that $p_0 > 0$ is equivalent to $x < 1 - \frac{d}{a}$ in this parameter region. Furthermore given that $x < 1 - \frac{d}{a}$ stability of $p = p_1$ requires

$$\begin{aligned} f'(p)_{|p=p_1} \quad &= \quad -\alpha \frac{[(a-d) - xa][x(1-d) - (a-d)]}{1-d-a} < 0 \\ &\Leftrightarrow \quad x > \frac{a-d}{1-d}. \end{aligned}$$

By noting that $1-\frac{d}{a} > \frac{a-d}{1-d}$ and $\frac{1-w-d}{1-d} \geq \frac{a-d}{1-d} \forall w \in [d, 1-a]$ the four cases from the proposition can be easily verified. ■

Appendix B (Endogenous Norm-strength)

In order to state the proof for Proposition 6, first note that $p = 0 \Rightarrow s = 1-x$ and $p = 1 \Rightarrow s = 1$. Denote

$$\frac{(1-w(s)-a)-x(1-d-a)}{x(a+d-1)} =: \Gamma(p)$$

and $\widetilde{p}$ the solution to $\Gamma(p) = p$ with corresponding norm strength $w(\widetilde{s})$.The following Lemma shows the existence of such a solution:

Lemma 1 *There exists $\widehat{x} \in [0,1]$ such that if $x \geq \widehat{x}$ there is a unique fixed point $\widetilde{p}$ (solving $\Gamma(p) = p$) with corresponding norm strength $w(\widetilde{s}) \in [1-a, d]$.*

Proof. First note that as $w(s) \in C^2, w(0) = 0$ and $w(1) = 1$ there exists $\widetilde{s}$ such that $w(\widetilde{s}) \in [1-a, d]$. Assume that $x \geq \frac{a+w(\widetilde{s})-1}{a+d-1} =: \widehat{x} \in [0,1]$. Furthermore note that $w(s) \in [1-a, d]$ implies $p \in [1-\frac{1-w^{-1}(1-a)}{x}, 1-\frac{1-w^{-1}(d)}{x}]$. Define

$$\Psi(p) = \Gamma(p) - p.$$

Obviously $\Psi(p)$ is a continous function of p. If $x \geq \widehat{x}$ we have that $\Psi(p)$ maps the non-empty, compact and convex interval $[1-\frac{1-w^{-1}(1-a)}{x}, 1-\frac{1-w^{-1}(d)}{x}]$ into $\mathbb{R}$. Furthermore $\Psi(\ 1-\frac{1-w^{-1}(1-a)}{x}) = \frac{1-w^{-1}(1-a)}{x} > 0$, and $\Psi(\ 1-\frac{1-w^{-1}(d)}{x}) = \frac{-w^{-1}(d)}{x} \leq 0$ if $x \geq \widehat{x}$. Consequently $\exists p^* \in [1-\frac{1-w^{-1}(1-a)}{x}, 1-\frac{1-w^{-1}(d)}{x}]$ such that $\Psi(p) = 0$. Uniqueness can be seen by noting that

$$\Psi'(p) = \frac{-w'(s)}{(a+d-1)} - 1 < 0$$

i.e. that $\Psi(p)$ is strictly decreasing. ■

Proof of Proposition 6

Proof. From Propositions 1 and 3 it follows that given $a+d > 1$ the interior zero p_2 will always be unstable independently of the strength of the norm. Note also that $a+d > 1 \Leftrightarrow 1-d/a < (a-d)/(1-d)$. Next examine the stability of the three candidates $p = 0$, $p = 1$ and $p = \widetilde{p}$. Focus first on the case where $p = 0$. Then we have that if $x > 1-w^{-1}(1-a)$, (2) corresponds to a prisoners' dilemma payoff-matrix and consequently $p = 0$ is unstable. If $x \in [1-w^{-1}(d), 1-w^{-1}(1-a)]$, (2) represents a stag-hunt game. In this case $p = 0$ is stable iff $x \in [1-\frac{d}{a}, \frac{a+w(\widetilde{s})-1}{a+d-1}]$. Finally if $x < 1-w^{-1}(d)$, cooperation is a dominant strategy in game (2). Remember that in this case $p = 0$ is locally stable iff $x > 1-\frac{d}{a}$. Noting that $\frac{a+w(\widetilde{s})-1}{a+d-1} > 0$ iff $x < 1-w^{-1}(1-a)$, we can summarize that $p^* = 0$ is locally stable iff

$$x \in [1-\frac{d}{a}, \frac{a+w(\widetilde{s})-1}{a+d-1}] \wedge \psi < \psi_1. \tag{16}$$

On the other hand $p = 1$ implies $w = w(1) = 1 > \max\{1 - a, d\}$. Then it is clear that $p^* = 1$ is locally stable iff

$$x < \frac{a-d}{1-d} \vee \left\{ x > \frac{a-d}{1-d} \wedge \psi > \psi_2 \right\}. \tag{17}$$

Finally noting that $\frac{a+w(\widetilde{s})-1}{a+d-1} < 1 \Leftrightarrow x > 1 - w^{-1}(d)$, we have for $\psi \to 0$ that $\widetilde{p}$ is locally stable iff

$$x > \frac{a + w(\widetilde{s}) - 1}{a + d - 1}. \tag{18}$$

Comparing conditions (16), (17) and (18) the five cases from the proposition follow. ■

Proof of Proposition 7

Proof. Observe first that $(a - d)/(1 - d) < 1 - \frac{d}{a}$ in this parameter region. Consider the equilibrium $p = 0$. In this case whenever $x > 1 - w^{-1}(d)$, (2) corresponds to a prisoners' dilemma payoff-matrix and consequently $p = 0$ is unstable. If $x \in [1 - w^{-1}(1 - a), 1 - w^{-1}(d)]$, (2) represents a chicken game. In this case $p = 0$ is stable iff $x > 1 - \frac{d}{a}$. Finally if $x < 1 - w^{-1}(1 - a)$, cooperation is a dominant strategy in game (2). Remember that in this case $p = 0$ is locally stable iff $x > 1 - \frac{d}{a}$.Summarizing thus $p^* = 0$ is locally stable iff

$$\psi < \psi_1 \wedge x \in [1 - \frac{d}{a}, 1 - w^{-1}(d)]. \tag{19}$$

By contrast $p = 1$ is locally stable iff

$$x < \frac{a-d}{1-d} \vee \left\{ x > \frac{a-d}{1-d} \wedge \psi > \psi_2 \right\}. \tag{20}$$

Observe then that in case (i) $p = 1$ is globally stable; in case (ii), $p = 1$ and $p = 0$ are unstable and $\forall p \in [0, 1]$ the norm is either strict or intermediate. Consequently $p = p_1$ is globally stable (Proposition 1); in case (iii) $p = 0$ is globally stable (as $1 - w^{-1}(d) > x > 1 - \frac{d}{a} > \frac{a-d}{1-d}$). In case (iv) $p = 0$ and $p = 1$ are unstable (as $x > \frac{a-d}{1-d}$). We have that $\forall p$ with $w(s) > d$, $\dot{p} < 0$. Whereas $\forall p$ such that $w(s) < d$, $\dot{p} > 0$.The globally stable equilibrium is thus the polymorphic state where the norm switches from being weak to being intermediate. This is the state where $w(s) = d$ or equivalently where $p = \widehat{p}$. ■

Statement of result Case $\mathbf{a + d = 1}$

Whenever $w < 1-a$ defection is a dominant strategy for both types, whereas whenever $w > 1-a$ defection is a dominant strategy for a 0-type and cooperation for a w-type.[30] If $\psi \to 0$ and

(i) $0 < x < \frac{a-d}{1-d} (= 1 - \frac{d}{a})$, the globally stable equilibrium is $p^* = 1$.

(ii) $x \in [\frac{a-d}{1-d}, 1 - w^{-1}(d)]$, the globally stable equilibrium is $p^* = 0$.

(iii) $x > 1 - w^{-1}(d)$, the globally stable equilibrium is $p^* = 1 - \frac{1-w^{-1}(1-a)}{x}$.

[30] If $w = 1 - a = d$ the bilateral game represented by A^w is trivial as all payoffs (matrix-entries) are equal.